出現頻度と連接頻度に基づく専門用語抽出

湯本紘彰*1、 森辰則*2、中川裕志*3
*1,2 横浜国立大学 *3 東京大学
E-mail: {hir,mori}@forest.dnj.ynu.ac.jp nakagawa@r.dl.itc.u-tokyo.ac.jp

本論文では，専門用語を専門分野コーパスから自動抽出する方法の提案と実験的評価を報告する。本論文では名詞 (単名詞と複合名詞) を対象として専門用語抽出について検討する。基本的アイデアは、単名詞のバイグラムから得られる単名詞の統計量を利用するという点である。より具体的に言えば、ある単名詞が複合名詞を形成するために連接する名詞の頻度を用いる。この頻度を利用した数種類の複合名詞スコア付け法を提案する。NTCIR1 TMREC テストコレクションによって提案方法を実験的に評価した。この結果、スコアの上位の 1,400 用語候補以内においては, 単名詞バイグラムの統計に基づく提案手法が優れていた。

Term Extraction Based on Occurrence and Concatenation Frequency

Hiroaki Yumoto† Tatsunori Mori† Hiroshi Nakagawa††
†Yokohama National University ††University of Tokyo

In this paper, we propose a new idea of automatically recognizing domain specific terms from monolingual corpus. The majority of domain specific terms are compound nouns that we aim at extracting. Our idea is based on single-noun statistic calculated with single-noun bigrams. Namely we focus on how many nouns adjoin the noun in question to form compound nouns. In adition, we combine this measure and frequency of each compound nouns and single-nouns, which we call FLR method. We experimentally evaluate these methods on NTCIR1 TMREC test collection. As the results, when we take into account up to 1,400 highest term candidates, FLR method performs best.

1 はじめに

自動用語抽出は専門分野のコーパスから専門用語を自動的に抽出する技術として位置付けられる。従来、専門用語の抽出は専門家の人手によらねばならず、大変に人手と時間がかかるため up-to-date な用語辞書が作れないという問題があった。それを自動化することは意義深いことである。専門用語の多くは複合語、とりわけ複合名詞であることが多い。よって、本論文では名詞 (単名詞と複合名詞) を対象として専門用語抽出について検討する。筆者らが専門分野の技術マニュアル文書を解析した経験では多数を占める複合名詞の専門用語は少数の基本的なかつこれ以上分割不可能な名詞 (これを以後、単名詞と呼ぶ) を組み合わせて形成されている。この状況では当然、複合名詞とその要素である単名詞の関係に着目することになる。

専門用語のもうひとつの重要な性質として [KU96] によれば、ターム性があげられる。ターム性とは、ある言語的単位の持つ分野固有の概念への関連性の強さである。当然、ターム性は専門文書を書いた専門家の概念に直結していると考えられる。したがって、ターム性をできるだけ直接的に反映する用語抽出法が望まれる。

これらの状況を考慮すると、以下のような理由により複合名詞の構造はターム性と深く関係してくることが分かる。ターム性は通常 tf×idf のような統計量で近似されるが、tf×idf といえども表層表現を利用した近似表現に過ぎない。やはり書き手の持っている概念を直接には表していない。第二の理由は、単名詞:N が対象分野の重要な概念を表しているなら、書き手は N を単独で頻繁に使うのみならず、新規な概念を表す表現として N を含む複合名詞を作りだすことも多いことである。

このような理由により、複合名詞と単名詞の関係を利用する用語抽出法の検討が重要であることが理解できる。この方向での初期の研究に [CP95] があり、英語、フランス語のコーパスから用語抽出を試みているが、テストコレクションを用いた精密な評価は報告されていない。中川ら [HT98] は、この関係についてのより形式的な扱いを試みている。そこでは、単名詞の前あるいは後に連接して複合名詞を形成する単名詞の種類数を使った複合名詞の重要度スコア付けを提案していた。この考え方自体は [Fun95] が非並行 2 言語コーパスから対訳を抽出するとき用いた context heterogeneity にも共通する。その後、中川らはこのスコア付け方法による用語抽出システムによって NTCIR1 の TMREC(用語抽出) タスク

に参加し良好な結果を出してはいる。が、彼らの方法はある単名詞に連接して複合名詞を構成する単名詞の統計的分布を利用する方法の一実現例に過ぎない。本論文では彼らの考え方を一般化する。その一般化した枠組の中でより最適な方法を模索することによって、彼らのアイデアの本質的特徴を明かにする。また、もうひとつの有力な用語抽出法である C-value による方法 [FA96] との比較を通じて、提案する方法により抽出される用語の性質などを調べる。

以下、2 節では用語抽出技術の背景、3 節では単名詞の連接統計情報を一般化した枠組、4 節では NTCIR1 TMREC のテストコレクションを用いての実験と評価について述べる。

2 用語抽出技術の背景

単言語コーパスからの用語抽出は少なくとも二つのフェーズがある。第一フェーズは、用語の候補の抽出である。第二フェーズは第一フェーズで抽出された候補に用語としての適切さを表すスコア付けないし順位付けである。この後に順位付けられた用語候補集合の中から適切な数の候補を用語として認定するという第三のフェーズがある。しかし、第三フェーズは認定したい用語数の設定など外部的要因に依存するところもあるので、本論文ではその技術的詳細に立ち入らないことにする。

2.1 候補抽出

西欧の言語と異なって空白のような明確な語境界がない日本語や中国語では、情報検索に使う索引語として文字 N-gram も考えられる [FC93, LWW97]。しかし、専門用語という観点に立てばやはり人間に理解できる言語単位でなければならず、結果として単語を候補にせざるをえない。また、NTCIR1 TMREC で使用されたテストコレクションでも単語を対象にしている。さて、単語も内実は単名詞と複合語に分かれる。関連する過去の研究では単語よりは複雑な構造である連語 (Collocation) や名詞句の抽出を目標にする研究 [SM90, Sma93, FA96, HN96, SSN97] が多い。連語や複合語のような言語単位を対象にする場合には、それらはより基本的な構造から構成されることを仮定しなければならない。ここでは、単名詞を最も基本的な要素とする。用語候補が単名詞のどのような文法的構造によって構成されるかという問題も多く研究されてきた [Ana94]。どのような構造を抽出するにせよ、まずコーパスの各文から形態素解析によって単語を切り出す必要がある。形態素解析の結果としては各単語に品詞タグが付けられる。よって、複合名詞を抽出するなら、連続する名詞を抽出すればよい。これまでの研究では、名詞句、複合名詞 [HN96, T00, HT98]、連語 [SM90, DGL94, FA96, SSN97] などを抽出することが試みられた。

2.2 スコア付け

前節で述べた用語候補抽出の後、用語候補に用語としての重要度を反映するスコア付けを行う。当然ながら、用語としての重要度はターム性を直接反映すると考えてよく、それゆえにスコアはターム性を反映したものが望ましい。しかし、ターム性というのは前にも述べたように直接計算することが難しい。このため、tf×idf のような用語候補のコーパスでの頻度統計で近似することがひとつの方法である。一方、[KU96] は用語の持つべきもうひとつの重要な性質、ユニット性を提案している。ユニット性とは、ある言語単位 (例えば、連語、複合語など) がコーパス中で安定して使用される度合いを表す。これを利用するスコアも用語の重要度を表す有力な方法である。例えば、Ananiadou らが [FA96, FA99] で提案している C-value は入れ子構造を持つコロケーションからユニット性の高い要素に高いスコアを付ける有力な方法である。[T00] は、分野固有の用語が分野に固有でない一般の用語からどのように離れたところに分布するかをもってターム性を計ろうとしている。[KA00] は日英 2 言語コーパスを用い、日本語の用語の対訳が英語のコーパスの対応する部分にも共起することがターム性を表わすというアイデアに基づいた用語抽出法を提案している。同様の考えは [DGL94] にも見られる。これらの研究は、用語の現れ方や使用統計に基礎をおくものである。一方、[HT98] は、単名詞と複合語の関係という用語の構造に着目してターム性を表わそうとしている。本論文の次節以降で我々は、ターム性を直接的に捉えようとする [HT98] を発展させようと試みる。

3 単名詞の連接統計情報の一般化

3.1 単名詞の連接

2 節の用語抽出技術の背景で述べた多くの研究では実質的に用語の対象にしているのは名詞である。実際、専門用語の辞典に収録されている用語も大多数は名詞である。例えば、[平山 96, 長尾 90, 青木 93] などでは収録されているのはほとんどが名詞である。そこで本研究では単名詞のうちでも名詞、そして複合名詞だけを対象にした。実際、用語の大多数は [平山 96, 長尾 90, 青木 93] に見られるように複合名詞である。しかし、これらの複合名詞の要素となる単名詞はあまり多数にのぼるわけではない。この考え方から、単名詞に連接して複合名詞を構成する単名詞の異なり数に着目するというアイデア [HT98] が生まれる。しかし、連接する単名詞の異なり数だけではなく、頻度など他の要素も考慮する方向での一般化を行いうことは重要である。連接する単名詞

のどのような性質に着目したときに性能の良いスコアになるかを調べるのが本論文の課題のひとつである。

まず、特定のコーパスを想定したとき、単名詞:Nが連接する状況すなわち語基バイグラムを一般的に図 1 のように表わす。

$$\begin{array}{llll} [LN_1 & N](\#L_1) & [N & RN_1](\#R_1) \\ [LN_2 & N](\#L_2) & [N & RN_2](\#R_2) \\ : & & : & \\ [LN_n & N](\#L_n) & [N & RN_m](\#R_m) \end{array}$$

図 1: 単名詞 N を含む単名詞バイグラムと左右連接単名詞の頻度

図 1 において、LN_i $(i = 1, ..., n)$ は、単名詞バイグラム $[LN_i \quad N]$ の N の左方に連接する n 種類の単名詞を表わし、単名詞バイグラム $[N \quad RN_i]$ の RN_i $(i = 1, ..., m)$ は N の右方に連接する m 種類の単名詞を表わす。また、() 内の $\#L_i$ $(i = 1, ..., n)$ は N の左方に連接する n 個の単名詞の頻度を表わし、$\#R_i$ $(i = 1, ..., m)$ は N の右方に連接する m 個の単名詞の頻度を表わす。もちろん、単名詞バイグラム $[LN_i \quad N]$ や $[N \quad RN_j]$ はより長い複合名詞の一部分であってもよい。以下にコーパスから抽出した “トライグラム”という単名詞を含む用語候補集合の簡単な作例を示す。

例 1: 単名詞バイグラム
トライグラム 統計、単語 トライグラム、クラス トライグラム、単語 トライグラム、トライグラム 抽出、単語 トライグラム 統計、文字 トライグラム

この例を図 1 に示す形式で表記すると以下のようになる。

単語 トライグラム (3)　　トライグラム 統計 (2)
クラス トライグラム (1)　トライグラム 抽出 (1)
文字 トライグラム (1)

図 2: 単名詞 “トライグラム”を含む単名詞バイグラムと左右連接単名詞の頻度の例

3.2 単名詞バイグラムを用いるスコア

単名詞バイグラムには異なり数の他に頻度情報 $\#L_i, \#R_j$ がある。この二つの要因を組み合わせ方としては種々の方法が考えられるが、簡単なのは異なり単名詞毎の頻度に何らかの関数を施して総和をとる方法であり、次式で表わされる。ただし、記法は図 1 の記号を用いる。

$$LN(N) = \sum_{i=1}^{n}(\#L_i) \quad (1)$$

$$RN(N) = \sum_{i=1}^{m}(\#R_i) \quad (2)$$

$\#LN(N), \#RN(N)$ は、それぞれ N の左方、右方に連接して複合名詞を形成する全単名詞に頻度である。図 2 の例だと、$\#LN($トライグラム$) = 5$、$\#RN($トライグラム$) = 3$ である。また、ここで連接頻度の情報を大きく反映させることもできるが実験においては大差が見られなかった。

3.3 複合名詞への拡張

以上のような方法で単名詞 (単名詞) のスコア付けができた。しかし、我々が注目している用語は単名詞だけではなく、複数の単名詞から生成される複合名詞も含まれる。先に述べたように専門用語ではむしろ複合名詞が多数であるので、単名詞のスコアを複合名詞に拡張することが必要である。複合名詞のスコア付けには、ふたつの考え方がある。第一の考え方は、複合名詞のスコアはその構成単名詞数すなわち長さに依存するというものである。この考え方に従えば、長い複合名詞ほど高いスコアがつくことが自然である。第二の考え方は、スコアは複合名詞の長さに依存しないというものである。この考え方に従えば、長さに対して依存しないような正規化が必要になる。専門用語に複合名詞が多いことは認めるにしても、長い程、あるいは逆に短い程、重要であるという根拠は今のところない。よって、我々は第二の考え方を採る。

まず、単名詞 $N_1, N_2, ..., N_L$ がこの順で連接した複合名詞を CN とする。CN のスコアとして各単名詞のスコアの平均をとれば、我々の採った第二の考えに沿った CN の長さに依存しないスコアを定義できる。ここでは、相乗平均を採用する。ただし、CN の構成要素の単名詞のスコアが一つでも 0 になると CN のスコアが 0 になってしまうのを避けるために次式で CN のスコア: $LR(CN)$ を定義する。

$$LR(CN) = (\prod_{i=1}^{L}(LN(N_i)+1)(LR(N_i)+1))^{\frac{1}{2L}} \quad (3)$$

例えば、図 2 の場合、$LR($トライグラム$) = \sqrt{(3+1)(5+1)} = 4.90$ である。(3) では CN の長さ L の逆数でべき乗しているので、$LR(CN)$ は CN の長さに依存しないようになる。したがって、単名詞も複合名詞も同じ基準でそのスコアを比較できる。なお、ここで定義した相乗平均の他に相加平均を用いる方法もあるが、以下では実験において若干性能の良かった相乗平均のみについて議論する。

3.4 候補語の出現頻度情報

(3) の $LR(CN)$ は単名詞、複合名詞の出現頻度における情報を考慮しなかった。そこで、候補語がが独立で使用された場合の頻度 $f(N)$ を考慮すべく、(3) を次のように補正した $FLR(CN)$ を定義する。

$$FLR(CN) = f(CN) \times LR(CN) \quad (4)$$

$f(CN)$ は候補語 CN が単独で出現した頻度である。例えば、例 1 の場合、“トライグラム”は 3 回独立に出現した単名詞なので、FLR(トライグラム$) = 3 \times \sqrt{(3+1)(5+1)} = 14.70$ となる。

3.5 MC-value

比較のために、単名詞バイグラムによらない用語スコア付けとして C-value [FA96] を考える。C-value は次式で定義される。

$$\text{C-value}(a) = (length(a) - 1) \times (n(a) - \frac{t(a)}{c(a)}) \quad (5)$$

ここで、a は複合名詞[1]、
$length(a)$ は a の長さ (構成単名詞数)、
$n(a)$ はコーパスにおける a の出現回数
$t(a)$ は a を含むより長い複合名詞の出現回数
$c(a)$ は a を含むより長い複合名詞の異なり数
である。ところがこの式だと、$length(a) = 1$ すなわち a が単名詞の場合 C-value が 0 になってしまい、適切なスコアにならない。C-value 以前の類似の方法の [Kit94] では、複合語を認識するための計算コストを用語の重要度評価に用いていた。C-value においても、このような背景から、一度複合名詞が切り出された後は、その構成要素の名詞数に比例する認識コストが重要度になる。ただし、複合名詞全体がすでに認識されている場合、名詞を順に認識していけば、最後の名詞を認識する手間は必要なくなる。したがって、(5) では $(length(a) - 1)$ となる。しかしながら、人間が言葉を認識する上では全ての構成要素の単名詞を認識していると考えられる。そこで、我々は [FA96] の定義を次のように変更した。また、変更した定義を以後、MC-value と呼ぶ。

$$\text{MC-value}(a) = length(a) \times (n(a) - \frac{t(a)}{c(a)}) \quad (6)$$

例 1 の場合、MC-value(トライグラム)=(7-7/5)=5.6 である。

[1] [FA96] では nested collocation

4 実験および評価

4.1 実験環境および方法

本節では、まず実験の主な環境となるテストコレクションについて述べる。我々が用いたのは NTCIR-1 の TMREC タスクで用いたテストコレクションである [kag99]。1999 年に行われた NTCIR-1 のタスクのひとつであった TMREC では、日本語のコーパスを配布して用語抽出を行う課題が行われた。主催者側が人手で準備した正解用語に対して参加システムが抽出した用語の一致する度合いを評価した。日本語コーパスは、NACSIS 学術会議データベースから収録された 1,870 の抄録からなる。対象の分野は、情報処理である。主催者側で準備した正解用語は 8,834 語であり、単名詞と複合名詞が多く含まれる。参加システム側で形態素解析を行うタスクと、主催者側で予め行って形態素解析済みコーパスを配布して利用するタスクがあった。我々は、形態素解析済みで品詞タグ付きのコーパスを利用した。

我々は、この品詞タグ付きのコーパスから用語候補として連続する名詞を抽出した。ただし、”的”と”性”でおわる形容詞は分野固有の複合語の用語に含まれることが多いと考え、例外として単名詞扱いしている。この結果、用語候補数は 16,708 になった。これらを 3 節に述べた諸方法でスコア付けし、スコアの高い用語候補から順にソートする。こうして作られた用語候補を上位から PN 個取り出した場合について、NTCIR-1 TMREC テストコレクションとして供給された正解用語とつきあわせて、抽出正解用語数、適合率を計算し評価する。これらは次式で定義される。

$$\text{抽出正解語数 } (PN) = \text{上位 } PN \text{ 候補中の正解語数} \quad (7)$$

$$\text{適合率 } (PN) = \frac{\text{抽出正解用語数 } (PN)}{PN} \quad (8)$$

なお、候補語を、上位 3,000 語まで 100 語ごとに評価した。以下の節で、この実験結果を示し、評価を行う。

4.2 各方法の比較実験および考察

ここでは、1) 連接頻度 $LN(N), RN(N)$ を用いた LR、2) LR に候補語の独立出現数を考慮した FLR、3) MC-value、を用いたスコアで順位付けした場合について、PN が上位 3,000 語まで、100 語ごとに評価する。

まず、図 3 に LR の手法によって抽出された候補語 3,000 語のうち、正解語との完全一致数と、正解語を含んだ候補語もカウントした部分一致数を示す。正解語を含んだ候補語も正解とすると 3,000 語

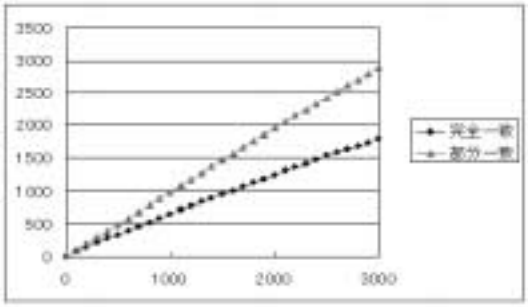

図 3: LR による候補語上位 3,000 語における完全一致数と部分一致数

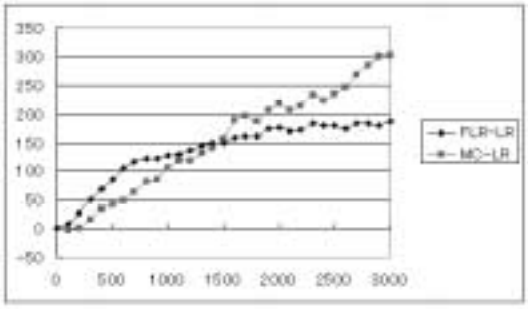

図 4: LR との差をとった FLR,MC-value における完全一致数

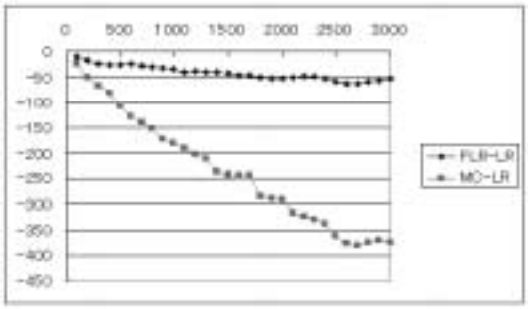

図 5: LR との差をとった FLR,MC-value における部分一致数

までではかなりの部分をカバーしていることがわかる。これに対して、FLR、MC-value を比較する。$\#LN(N), \#RN(N)$ の完全一致数、部分一致数の差をとった図を 4,5 に示す。図 4,5 では、LR の抽出用語数から、FLR,MC-value の抽出用語数をそれぞれ引いた用語数を示している。完全一致数では FLR、MC-value 共に LR を上回る結果となった。さらに、1,400 語までは FLR が最も優れた結果を示し、それ以降は MC-value がこれを上回った。また、部分一致数では FLR,MC-value 共に LR を下回る結果となった。しかしながら、FLR と LR の手法は大差はないが、MC-value はこれらを大きく下回る結果となった。これらを見てもわかるように、我々の提案する手法は完全に間違った候補語は抽出されにくいが、MC-value は正解語とまったく関係のない候補語も抽出される傾向にあるといえる。

4.3 抽出用語の性質

さて、これまでは抽出用語の質をそのまま候補語中の正解用語数で議論してきた。しかし、テストコレクションの正解が実用的にどのくらい有効な指標になっているかは議論の余地がある。そこで抽出用語に対する直接的な評価を以下に試みる。まず、用語の長さは抽出用語の品質に密接に関係する。そこで、各スコア付けの方法における上位から並べた正解用語の長さを図 6 に示す。ただし、長さは複合名詞を構成する単名詞数で表わすことにした。

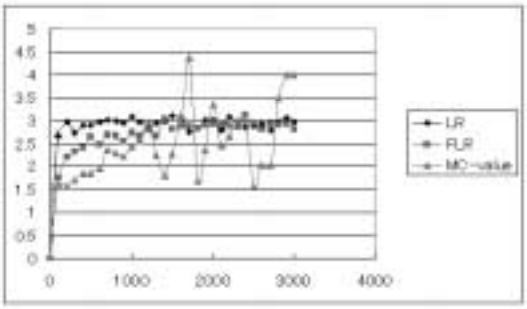

図 6: 各手法における 100 語毎の平均語長

図 6 を見ると、候補語上位 1,400 語付近までで MC-value 方法は他の方法に比べて平均語長が平均的に短い。MC-value では語長の短い語が高いスコアを得る傾向にある。ところが、上位 1,400 語までは FLR が最も多くの正解用語を抽出している。上位 1,400 語以降、MC-value は語長の長い語も抽出するようになるにつれて、より多くの正解用語を抽出するようになる結果となった。

次に具体的な抽出用語例を示そう。全てを示すことはスペースの関係でできないが、最上位 15 語の抽出用語を示して各スコア付けの特徴について考えてみる。

表 1: スコアの最上位の 15 用語候補

無印:正解用語 ×:不正解語

LR		FLR		MC-value	
知識		知識		学習者	
学習知識		システム		問題解決	
学習		問題		システム	
言語的知識		学習		知識	
知識システム		モデル		研究	×
学習システム		情報		本稿	×
問題知識	×	問題解決		手法	×
学習問題		設計		問題	
言語的		知識ベース		知識ベース	
システム		推論		論文	×
問題		支援	×	方法	×
論理的知識		知識表現		支援システム	
学習支援システム		エージェント		計算機	
設計知識		学習者モデル		情報	
学習問題解決システム		構造	×	モデル	

表 1 に各手法におけるスコアの最上位 15 候補を示す。この結果を見ると、明らかに LR によるスコア付の上位候補は複合名詞が多い。一方、FLR と MC-value のスコア付けの上位候補には単名詞が多い。FLR では、出現頻度の高い単名詞を優遇する補正をしているし、MC-value でも単名詞の頻度がその単名詞を含む複合名詞の頻度を強く反映した構造になっているから、この結果は偶然ではない。MC-value の場合、”研究、論文、方法、手法”などという分野の用語でない名詞が多く抽出されているが、これも大量かつ多種類の複合名詞に含まれるであろうこと、および MC-value が多数かつ多種類の複合名詞に含まれる単名詞のスコアを高くつけることから得られる帰結である。

4.4 NTCIR-1 の結果との比較

ここまで述べてきたスコア付け方法の客観的評価を行うために NTCIR-1 TMREC タスクに参加した上位の成績を残したチームとの比較する。なお、NTCIR-1 には C-value によるスコア付けをするチームも参加しているが、NTCIR-1 の参加規定によりどのチームかは不明である。しかし、後で述べるように本論文で提案した C-value を修正した MC-value が良好な結果を示していることから、我々の C-value の修正法には若干の独自性が認められると考えられる。NTCIR-1 TMREC の上位 2 チーム、以後 N1,N2 と呼ぶ。N1,N2 と本論文で性能の良かった FLR および MC-value に各スコア付け方法において、上位から 1,000 語毎、3,000 語までの範囲での適合率を表 2 に示す。

表 2: NTCIR-1 TMREC 参加上位 2 チームと FLR、MC-value の比較 (適合率)

PN	FLR	MC-value	N1	N2
1 から 1,000	.773	.754	.705	.744
1,001 から 2,000	.635	.707	.607	.584
2,001 から 3,000	.562	.640	.618	.518

表 2 によれば、スコア付け 1001〜2000,2001〜3000 語の部分では MC-value が他を上回ったが、1〜1000 語部分での抽出精度は我々の提案した FLR によるスコア付けが、最も優れた結果を示した。また、スペースの関係上、表には載せていないが、候補語数が多くなるにつれて、他の手法は適合率を急激に落とすが、FLR の抽出精度の下がり方はなだらかであった。このことは FLR が安定して正解語を抽出していることを示している。我々は、名詞の連続だけを取り出したが、正解語の中には形容詞と名詞の連接や、助詞”の”によってつながった用語もある。これらを広く抽出すれば再現率は高まるが、上位のスコアの抽出後においてすら非正解語を多数抽出してしまい、あまり好ましくないと言える。

5 おわりに

本論文では、専門分野コーパスからの専門用語の抽出法について検討した。まず、用語抽出技術の背景を述べ、次に本論文の核心である単名詞 N に連接する単名詞の頻度の統計量を利用する N のスコア付けを一般的に表わす枠組みを提案した。これらスコア付け方法を複合名詞のスコア付けに拡張した。また、比較として、既存の C-value を修正した MC-value について述べた。これらのスコア付け法を NTCIR-1 TMREC タスクのテストコレクションに適用して結果を評価した。この結果、スコア上位の候補においては我々の提案する FLR の性能が優れていた。より包括的に（1,500〜10,000 語）専門語を抽出したいのなら、MC-value のほうが優れた結果を示すが、正解語を含む長めの語でよいのであれば、FLR は大部分をカバーすることができる。今後の課題としては、より多様な情報例えば文脈情報を利用して用語抽出の性能の向上を計ることが重要である。しかし、一方で、専門分野の用語として真に欲しいのはどのような性質を持つ用語なのかを定式化するという根本的問題も考察していく必要があろう。このような考察は哲学的なものというよりは、実際のコーパスの統計処理を用いた実験的なものでなければ実用性に乏しい。その意味で、このような観点から設計した用語抽出タスクを企画することも望まれる時期にきているのではないだろうか。

6 謝辞

本研究は文部科学省の研究費補助金によって行われている。

参考文献

[Ana94] Sophia Ananiadou. A methodology for automatic term recognition. In *COLING'94*, pp. 1034 – 1038, 1994.

[CP95] Enguehard C. and L. Pantera. Automatic natoinal acquisition of terminology. *Journal of Quantative Linguistics*, Vol. 2, No. 1, pp. 27 – 32, 1995.

[DGL94] Beatrice Daille, Eric Gaussier, and Jean Marc Lange. Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of COLING'94*, pp. 515 – 521, 1994.

[FA96] Katerina T. Frantzi and Sophia Ananiadou. Extracting nested collocations. In *COLING'96*, pp. 41 – 46, 1996.

[FA99] Katerina T. Frantzi and Sophia Ananiadou. The c-value/nc-value method for atr. *Journal of NLP*, Vol. 6, No. 3, pp. 145 – 179, 1999.

[FC93] Hideo Fujii and W. Bruce Croft. A comparison of indexing techniques for japanese text retrieval. In *Proceedings of SIGIR'93*, pp. 237 – 246, 1993.

[Fun95] Pascale Fung. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In *Proceedings of the Workshop on Very Large Corpora*, pp. 173–183, 1995.

[HN96] Toru Hisamitasu and Yoshihiko Nitta. Analysis of japanese compound nouns by direct text scanning. In *Proceedings of the 15th International Conference on Computationa Linguistics*, pp. 550 – 555, 1996.

[HT98] Nakagawa. H and Mori. T. Nested collocation and compound noun for term recognition. In *Proceedings of the First Workshop on Computational Terminology COMPTERM'98*, pp. 64 – 70, 1998.

[KA00] Kageura K, Tsuji K and Aizawa A. Automatic thesaurus generation through multiple filtering. In *Proceedings of the 18th International Conference on Computational Linguistics*, pp. 397 – 403, 2000.

[kag99] Kyo kagura. TMREC task: Overview and evaluation. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp. 411 – 440, 1999.

[Kit94] K Kita. A comparative study of auotomatic extraction of collocations from copora: Mutual infomation vs. cost criteria. *Journal of NLP*, Vol. 1, No. 1, pp. 21 – 29, 1994.

[KU96] Kageura. K. and B. Umino. Methods of automatic term recognition:a review. *Terminology*, Vol. 3, No. 2, pp. 259 – 289, 1996.

[LWW97] Wai Lam, Chi-Yin Wong, and Kam-Fai Wong. Performance evaluation of

character-, word- and n-gram-based indexing for chinese text retrieval. In *Proceedings of the and International Workshop on Information Retrieval With Asian Languages*, pp. 68 – 80, 1997.

[SM90] Frank A. Smadja and Kathleen R. Mckeown. Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th ACL*, pp. 252 – 259, 1990.

[Sma93] Frank Smadja. Retrieving collocations from text:xtract. *Computational Linguistics*, Vol. 19, No. 1, pp. 143 – 177, 1993.

[SSN97] Sayori Shimohata, Toshiyuki Sugio, and Junji Nagata. Retrieving collocations by co-occurences and word order constraints. In *Proceedings of the 35th ACL-8th EACL*, pp. 476 – 481, 1997.

[T00] Hisamitsu. T. A method of measuring term representativeness. In *Proceedings of 18th International Conference on Computational Linguistics*, pp. 320 – 326, 2000.

[青木 93] 青木繁 (編). 建築大辞典. 彰国社, 1993.

[長尾 90] 長尾真 (編). 岩波情報科学辞典. 岩波書店, 1990.

[平山 96] 平山博, 氏家理央 (編). 電子情報通信英和・和英辞典. 共立出版, 1996.